Complex Networks

Teachers: M. Emmerich, D. Garlaschelli, F. den Hollander.
Written examination: 30 January 2019, 10:00-13:00.

Open book exam: the lecture notes may be consulted, but no other material.

Answer each question on a separate sheet. Put your name, student number and the number of the question you are answering on each and every sheet. Provide full explanations with each of the answers!

Each question is weighted by a number of points, as indicated. The total number of points is 100. The final grade will be calculated as a weighted average: 30% for the homework assignments and 70% for the exam.

Success!

1. a. [**3 points**] The WWW has 50-55 billion webpages. Still, it is possible to navigate to almost every webpage within a few clicks only. Explain why this is so. What is this property called?

   b. [**2 points**] Explain why it is reasonable to model real-world networks as large random graphs.

   c. [**2 points**] Why is it useful to consider sequences of random graphs labelled by the number of vertices?

2. Consider the configuration model with n vertices with degrees $D_1, \dots, D_n$ that are drawn independently from a probability distribution f on $\mathbb{N}$ that is non-deterministic.

   a. [**3 points**] Compute the average number of neighbours ν of a vertex, say vertex 1.

   b. [**3 points**] Compute the average number of neighbours $\bar{\nu}$ of a randomly chosen neighbour of vertex 1.

   c. [**3 points**] Show that $\bar{\nu} > \nu$, and explain why this is so.

3. Consider the preferential attachment model with parameters $m = 1$ and $\delta = -\frac{1}{2}$.

   a. [**4 points**] List the possible outcomes of $\mathrm{PA}_1(1, -\frac{1}{2})$ and $\mathrm{PA}_2(1, -\frac{1}{2})$.

   b. [**5 points**] Compute the probability of each of these outcomes. Explain your answer.

   c. [**4 points**] Is the sequence $(\mathrm{PA}_n(1, -\frac{1}{2}))_{n \in \mathbb{N}}$ scale free?

4. Let $\mathbf{G}^*$ denote a simple undirected graph consistent with the degree se-

quence $\vec{k} = (1, 2, 4, 2, 1)$.

a. [**2 points**] Draw your graph and write the corresponding adjacency matrix.

b. [**2 points**] Now consider the Erdős-Rényi (ER) random graph with a fixed number $n = 5$ nodes and connection probability p as a model for $\mathbf{G}^*$. Find the value p^* that maximises the likelihood of generating $\mathbf{G}^*$.

c. [**3 points**] Calculate the expected value $\langle k_i \rangle$ of the degree of each node i under the ER model with $n = 5$ and $p = p^*$. For each vertex, determine whether the realized degree in $\mathbf{G}^*$ is larger or smaller than the corresponding expected degree in the ER model.

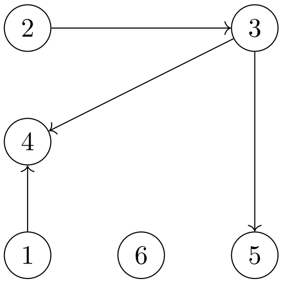
d. [**3 points**] Calculate the standard deviation σ_i of the degree of each verterx i under the ER model with $n = 5$ and $p = p^*$. Identify the vertices whose realized degree in $\mathbf{G}^*$ differs from the expected degree by more than one standard deviation (in either direction).

e. [**3 points**] Calculate the probability of occurrence of $\mathbf{G}^*$ under the ER model with $n = 5$ and $p = p^*$.

f. [**3 points**] Find the most probable and least probable graphs under the ER model with $n = 5$ and $p = p^*$. Compare these graphs with your graph $\mathbf{G}^*$.

5. Complex networks can be represented as an adjacency matrix, an adjacency list, and an edge list.

a. [**4 points**] Describe the graph in the picture as an adjacency matrix and as an igraph graph formula ($\hat{=}$ adjacency list). Use the syntax of igraph (R or Python).



b. [**2 points**] An ultradense graph is a graph where the degree of each node is greater than or equal to $n - k$ for some constant $k \ll n$ and n denoting the number of edges. How does the space complexity of an straightforward implementation of an adjacency matrix and of a adjacency list scale asymptotically in terms of n in the worst case? Use the Big O notation.

c. [**4 points**] Given a small and constant value of k and the knowledge that all data will be ultradense graphs with constant k. Suggest a data-structure that is space-efficient for storing an ultra-dense graph and provide its space efficiency.

6. Two graphs are isomorphic if one graph can be obtained from the other graph by re-assigning the node labels.

   a. [**3 points**] Given a graph with node set V and edge set E. How many isomorphic graphs do there exist?

   b. [**4 points**] Discuss an efficient way to uniformly at random generate a graph that is isomorphic to a given graph. Describe the algorithm by means of pseudo-code and determine its time complexity in the Big O notation.

   c. [**2 points**] Discuss briefly how sparsity (node degree bounded by a small constant k) can be exploited to check whether two graphs are isomorphic.

7. a. [**4 points**] Does the contact process on $\mathbb{Z}^d$ have a non-trivial threshold in any dimension $d \geq 1$?

   b. [**4 points**] Is the answer the same on a finite graph?

   c. [**6 points**] What is the most relevant quantity for an epidemiologist to know while monitoring the evolution of a virus spreading in a finite population? Explain your answer.

8. Given a generic binary undirected graph, let k_i denote the degree of vertex i and c_i the local clustering coefficient of vertex i. Imagine that you produce a scatter plot where each vertex i is represented as a point with coordinates (k_i, c_i) in the plane.

   a. [**4 points**] Describe what the scatter plot looks like in typical realizations of the Erdős-Rényi random graph model with n nodes and connection probability p. Explain your answer.

   b. [**4 points**] Describe what the scatter plot looks like in typical realizations of the canonical configuration model, as a function of the input degree sequence. Explain your answer.

   c. [**5 points**] Describe what the scatter plot looks like in different real-world networks, and what can be concluded from it about the clustering properties of these networks.

9. Consider a clique (fully connected, unconnected graph) with only three nodes: v_1, v_2, and v_3. Assume the SIS model with infection rate λ for all nodes and healing rate μ.

   9a. [**5 points**] Describe the generator matrix of the continuous-time Markov chain (CTMC) related to the SIS process. You can fill in the rates in the table below, where the bits represent network states, e.g., 101 means that v_1 and v_3 are infected and v_2 is in a susceptible state.

	000	100	010	001	110	101	011	111
000								
100								
010								
001								
110								
101								
011								
111								

9b. **[4 points]** Next, for simplicity, consider the SI process. Let node v_1 be just infected by a disease and the other nodes be in a susceptible state at some time t_0. What is the time that the system requires to reach the state where one additional node gets infected (on average). What is the probability of different states to get this state?

SOLUTIONS

1a. The reason is that WWW is ‘locally tree-like’, which makes distances small (of the order of the logarithm of the number of webpages). This property is called ‘small world’.

1b. Real-world networks consist of ‘nodes’ and ‘links’, which can be modelled as vertices and edges in a graph. Real-world networks are large and complex, which can be modelled with the help of randomness, because in general the full architecture of the network is not known.

1c. Sequences of random graphs allow us to model very large networks and pass to the limit of infinitely many vertices, in which case probabilistic limit theorems can be derived.

2a. $\nu = \sum_{k\in\mathbb{N}} kf(k)$, because $f(k)$ is the probability that vertex 1 has k neighbours.

2b. $\bar{\nu} = \sum_{k\in\mathbb{N}} k^2 f(k)/\sum_{k\in\mathbb{N}} kf(k)$, because $\bar{f}(k) = kf(k)/\sum_{k\in\mathbb{N}} kf(k)$ is the probability that a randomly chosen neighbour of vertex 1 has k neighbours.

2c. The Cauchy-Schwarz inequality gives $\bar{\nu} > \nu$, unless f puts unit weight on a single k. In words, since a vertex with a large degree is more likely to share an edge with vertex 1 than a vertex with a small degree, the neighbours of vertex 1 on average have more neighbours than vertex 1 itself.

3a. $P_1(1,-\frac{1}{2})$ consist of a one vertex v_1 with a self-loop. $P_2(1,-\frac{1}{2})$ consist of two vertices v_1 and v_2, either with a self-loop each or with a self-loop at v_1 and an edge between v_1 and v_2.

3b. The single possibility for $P_1(1,-\frac{1}{2})$ has probability 1. The two possibilities for $P_2(1,-\frac{1}{2})$ have probability $\frac{1}{4}$, respectively, $\frac{3}{4}$.

3c. Yes. The probability for a vertex to have degree k decays like $k^{-\tau}$ as $k \to \infty$, with exponent $\tau = 3 + (\delta/m) = 2.5$.

4a. There is only one possible graph $\mathbf{G}^*$ consistent with the degree sequence $\vec{k} = (1,2,4,2,1)$, and its adjacency matrix is

$$\begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}.$$

4b. The value of p maximizing the likelihood for the ER with $n = 5$ and $p = p^*$ is

$$p^* = \frac{2L}{n(n-1)} = \frac{10}{5 \cdot 4} = \frac{1}{2}$$

where $L = 5$ is the number of vertices in $\mathbf{G}^*$.

4c. The expected degree $\langle k_i \rangle$ of each node i in the ER model with $n = 5$ and $p = p^*$ is

$$\langle k_i \rangle = p^*(n-1) = \frac{1}{2} \cdot 4 = 2$$

which is the same for all nodes. Therefore, in the graph $\mathbf{G}^*$, for nodes $i = 1, 5$ we have $k_i < \langle k_i \rangle$, while for nodes $i = 2, 4$ we have $k_i = \langle k_i \rangle$ and finally for node $i = 3$ we have $k_i > \langle k_i \rangle$.

4d. The standard deviation σ_i of the degree of each node i in the ER model with $n = 5$ and $p = p^*$ is

$$\sigma_i = \sqrt{p^*(1-p^*)(n-1)} = 1$$

which is again the same for all nodes. For each node i in graph $\mathbf{G}^*$, we have to check whether its realized degree k_i is within one standard deviation from the expected degree, i.e. whether $\langle k_i \rangle - \sigma_i \leq k_i \leq \langle k_i \rangle + \sigma_i$. Given the values $\langle k_i \rangle = 2$ and $\sigma_i = 1$ calculated above, this means checking whether $1 \leq k_i \leq 3$. We see that only node 3, which has degree $k_3 = 4$, is outside the above range.

4e. The probability of occurrence of $\mathbf{G}^*$ in the ER model with $n = 5$ and $p = p^*$ is given by

$$P(\mathbf{G}^*) = (p^*)^L (1-p^*)^{n(n-1)/2-L},$$

but since in this particular case $p^* = 1 - p^* = 1/2$, we have

$$P(\mathbf{G}^*) = (1/2)^{L+n(n-1)/2-L} = (1/2)^{n(n-1)/2} = 2^{-10}.$$

4f. Note that, since $p^* = 1/2$, for any other graph with $n = 5$ nodes we arrive at the same probability as calculated in the point above, because the result is independent of the number of links. This is due to the fact that $p^* = 1/2$ implies that both links (p^*) and 'non-links' ($1 - p^*$) occur with the same probability. Therefore, with $n = 5$ and $p = p^*$, in the ER *all graphs (whether 'similar' to* $\mathbf{G}^*$ *or not) occur with exactly the same probability*, equal to 2^{-10}.

5a.
```
g<-graph.adjacency
(matrix(c(0,0,0,1,0,0,
                0,0,1,0,0,0,
                0,0,0,1,1,0,
                0,0,0,0,0,0,
                0,0,0,0,0,0,
                0,0,0,0,0,0),nrow=6, ncol =6));

g<-graph.formula(1-+4,2-+3,3-+4:5,4,5,6)
```

5b. The space complexity is $\Theta(n^2)$ for the adjacency matrix and for the adjacency list.

5c. The space complexity can be reduced to $O(nk) = O(n)$ because, instead of storing the edges connected to a node in an adjacency list, one can store the edges in $\overline{E}$, i.e., the edges that are not represented.

6a. The number of isomorphic graphs is given by $n!$, since all the label permutations provide isomorphic graphs.

6b.
1. Initialize an array of all integers from $a[1] = 1, \ldots, a[|V|] = |V|$.
2. Shuffle the array by applying the Fisher-Yates method
3. For each edge (i, j) in E, output $(a[i], a[j])$.

The time complexity is $O(\max(|V|, |E|))$.

6c In order to check isomorphism by enumeration, only $O(k\lceil n/k\rceil) = O(nk^n)$ combinations need to be compared.

7a. Yes, $\lambda_d \in (0, \infty)$ for all $d \geq 1$. This result is obtained from three inequalities: $d\lambda_d \leq \lambda_1$, $2d\lambda_d \geq 1$, $\lambda_1 < \infty$.

7b. No, for any finite graph the critical threshold for an epidemic is ∞: eventually the whole population becomes healthy.

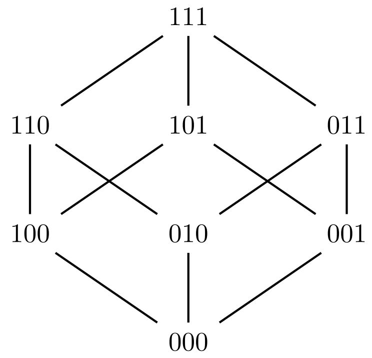
7c. What matters for an epidemic is whether the average time until the population becomes healthy is small or large. For large finite graphs that have certain regularity properties, there is a critical threshold above (below) which the average time grows exponentially (logarithmically) fast with the number of vertices.

8a. In typical realizations of the Erdős-Rényi model, the degrees of vertices are distributed around the expected value $\langle k\rangle = p(n-1) \approx pn$ according to a binomial distribution (which, in the sparse case, approaches a Poisson distribution) and the clustering coefficient of each node is narrowly distributed around its expected value $\langle c_i\rangle = p$ (which is the same for all noddes). This means that a typical scatter plot is a 'ball' of points randomly scattered around the point (pn, p).

8b. In typical realizations of the canonical configuration model with given degree sequence, the k_i coordinates of the scatter plot are fixed by the chosen degree sequence, while each of the c_i coordinates will be randomly scattered around its expected value $\langle c_i\rangle = \frac{\sum_{k\neq i,j}\sum_{j\neq i} p_{ij}p_{jk}p_{ki}}{\sum_{k\neq i,j}\sum_{j\neq i} p_{ij}p_{ki}}$, where $p_{ij} = x_i x_j/(1 + x_i x_j)$ and each x_i is such that $k_i = \sum_{j\neq i} p_{ij}$. If the input degree sequence is a constant vector, then the canonical configuration model reduces to the Erdős-Rényi model discussed in point 8a above. As the degree sequence becomes more and more heterogeneous, the scatter plot acquires a typically decreasing trend. If the degree distribution has a decreasing power-law tail, then the tail of the scatter plot is a decreasing power law as well.

8c. In real-world networks, the empirical scatter plot is generally decreasing, indicating a sort of hierarchical organization of triangles as a function of the degree of their participating nodes. Low-degree nodes are connected to a few high-degree nodes which are connected among themselves, thus closing many potential triangles and producing a large local clustering coefficient. On the other hand, high-degree nodes are connected to many nodes that have low degree and that are generally not connected among themselves, thus producting a low local clustering coefficient. In any case, what matters in order to assess whether there is a high or low level of clustering is not the empirical scatter plot itself, but rather its comparison with the one obtained in the configuration model with the same input degree sequence as the empirical one. High or low clustering is then signalled *locally* by some points of the empirical scatter plot being significantly above/below the points of the corresponding vertices in the scatter plot obtained in the configuration model. Both possibilities are observed in real-world networks. In some cases (e.g. interbank networks), an almost complete consistency with the configuration model is observed, indicating 'random' clustering.

9a. The state space has size 2^3, all possible settings of the three nodes to S and I. The generator matrix can be drawn as a graph.



	000	100	010	001	110	101	011	111
000	0	0	0	0	0	0	0	0
100	m	$-2l-m$	0	0	l	l	0	0
010	m	0	$-2l-m$	0	l	0	l	0
001	m	0	0	$-2l-m$	0	1	1	0
110	0	m	m	0	$-2m-l$	0	0	l
101	0	m	0	m	0	$-2m-l$	0	l
011	0	0	m	m	0	0	$-2m-l$	l
111	0	0	0	0	m	m	m	$-3m$

9b. The average time to infect the next node is $1/(2\lambda)$. The subsequent state is with equal probability ($p = 0.5$) the state 101 and 110.

9c. The total time can be simulated by the following program:

```
Time ← 0
For $i = 1$ to $N$
        Time ← Time + ExpDistribution($1/(i(N-i))$)
```

4. Return Time

Remarks:
- This makes use of the fact that the probability that in a single differential time step more than one node gets infected is zero.
- Each susceptible node at time step i can get infected from i nodes (at rate $i\lambda$), and there are $n - i$ nodes that can get infected in the next time step.